DOCUMENT RESUME

ED 387 521 TM 023 815

AUTHOR	Wainer, Howard
TITLE	Depicting Error. Program Statistics Research. Technical Report No. 95-2.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	ETS-RR-95-11
PUB DATE	Mar 95
CONTRACT	R999B40013
NOTE	46p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Access to Information; *Communication (Thought Transfer); Data Analysis; *Information Dissemination; National Surveys; *Tables (Data)
IDENTIFIERS	Data Display; *Error Detection; *Error Reports; National Assessment of Educational Progress

ABSTRACT

Communicating data without some measure of their precision can lead to misinterpretation and incorrect inferences. Several conventions for displaying error along with the data they modify are described and illustrated. Alternatives are offered that seem to provide improvements in the effective communication of error as well as increasing their ease, and hence their likelihood of use. These alternatives are illustrated principally with data from the National Assessment of Educational Progress. Sometimes, with tabular presentation, this involves calculating and displaying the appropriate error terms for the most likely uses of the data. In other situations, it means making the precision an integral part of the display. Two tables and 14 figures illustrate the discussion. (Contains 22 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 387 521

RR-95-11

Depicting Error

Howard Wainer
Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Coley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

PROGRAM
STATISTICS
RESEARCH

Technical Report No. 95-2

Educational Testing Service
Princeton, New Jersey 08541

BEST COPY AVAILABLE

TM023615

Depicting Error

Howard Wainer
Educational Testing Service

Program Statistics Research
Technical Report No. 95-2

Research Report No. 95-11

Educational Testing Service
Princeton, New Jersey 08541

March 1995

Copyright © 1995 by Educational Testing Service. All rights reserved.

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

Depicting error

Howard Wainer[§]
Educational Testing Service

Abstract

Communicating data without some measure of their precision can lead to misinterpretation and incorrect inferences. Several conventions for displaying error along with the data they modify are described and illustrated. Alternatives are offered that seem to provide improvements in the effective communication of error as well as increasing their ease, and hence their likelihood, of use. These alternatives are illustrated principally with data from the National Assessment of Educational Progress.

[§] This research was supported by contract number R999B40013 from the National Center for Education Statistics to the Educational Testing Service, Howard Wainer, Principal Investigator. I am pleased to be able to acknowledge this help. I am also grateful to Alan MacEachren, Eugene Johnson, Robert Mislevy, Charles P. Nesko, Linda Steinberg and John W. Tukey for their helpful comments on an earlier draft.

Depicting error

Introduction

"One hallmark of the statistically conscious investigator is his firm belief that however the survey, experiment, or observational program actually turned out, it could have turned out somewhat differently. Holding such a belief and taking appropriate actions make effective use of data possible. We need not always ask explicitly 'How much differently?' but we should be aware of such questions. Most of us find uncertainty uncomfortable...(but)... each of us who deals with the analysis and interpretation of data must learn to cope with uncertainty."

(Mosteller & Tukey, 1968, p. 100)

Whenever we discuss information we must also discuss its accuracy. Effective display of data must:

1. remind us that the data being displayed do contain some uncertainty, and then
2. characterize the size of that uncertainty as it pertains to the inferences we have in mind, and in so doing
3. help keep us from drawing incorrect conclusions through the lack of a full appreciation of the precision of our knowledge.

For the examples chosen here we will assume that we have estimates of precision and the task is strictly one of effective display. This assumption is admittedly a big one. Great efforts and much imagination have been expended in the search for the true uncertainty. Modern statistics has gone far beyond characterizing error by dividing the observed standard deviation by $\sqrt{n}$. Methods of sensitivity analysis through resampling or multiple imputation are some attempts to measure the real uncertainty. Despite our appreciation for the importance of this work, we will ignore how uncertainty is calculated and proceed assuming that, in whatever way was required, such estimates are available and our task is simply to convey them as well as we can.

Errors in Tables

Let us begin our examination of the depiction of error with a typical tabular display. Shown in Table 1 is a data table extracted from a much larger display. Omitted are 30 states and two territories as well as another variable (Percentage in each category) and its standard error. As abbreviated, this table still contains the elements important to our discussion without being unwieldy. The only deviation from standard format in this table is that the standard error of each data entry is in a separate column rather than denoted within parentheses adjacent to the entry they characterize. This was done primarily to ease text manipulation. There are three lines added to the bottom of this table. These lines are the point of this discussion.

Insert Table 1 About Here

Showing the standard errors of statistics in this way certainly satisfies requirement (1). The visual weight given to the numbers connoting error is the same as that which conveys the data. This depiction does not however allow us to summarize easily the structure of error. For example, we cannot tell if the size of the error is related to the proficiency, nor is it easy to see how much variation there is in the error over the various states or across the categories of parental education. The answers to these kinds of questions are important for requirement (2). In this case, errors across states are quite homogeneous and so displaying the maximum value of the standard error as a summary allows us to extract a handy, conservative value from the error terms provided. Using such a summary term (rather than say a mean or a median) will lower the likelihood of our declaring states different when there is reasonable evidence that they might not be. This raises the second part of requirement (2) "uncertainty as it pertains to the inferences we have in mind". What kinds of question are likely to be used to query Table 1?

When national data are reported broken down by states it is natural to assume that the disaggregated data are meant to be compared. Thus the standard errors are merely the building blocks that are required to be able to construct the standard error of the difference so that the statistical significance of the observed difference between state means can be ascertained. An upper bound on this standard error is obtained by multiplying the maximum standard error by the $\sqrt{2}$. This result is reported on the first line of the section at the bottom of the table labeled "Error terms for comparisons." It is a handy rule-of-thumb for anyone wishing to make a single comparison. Thus suppose, for some reason, we are interested in comparing the performance of Hawaii's 8th graders whose parents were college graduates (267), with their counterparts in Delaware (274). We note that their mean scores differ by 7. This is twice the maximum standard error of the difference of the means (3.5). We can thus conclude that this observed difference is statistically significant beyond the nominal ($\alpha = .05$) value. We could have done this comparison for any single pair that was of special interest to us.

Although the comparison of a particular pair of states may be an occasional use for a data table like this, we suspect that comparing one state (perhaps our own) with all others is a more likely use. To do this correctly we need to control the artifactual increases in the error rate due to making many comparisons. The most common way to assure that the type I error rate is controlled is boosting the size of the standard error sufficiently so that the overall error rate remains at the nominal (.05) level. In the original table there were 41 states and so comparing any one of these with each of the others would yield a total of 40 separate t-tests. The Bonferroni (1936) inequality (presented carefully, with many examples, in Miller, 1966) provides us with a conservative rule for combining the standard error of the difference with the 40 different t-tests to yield the ordinate associated with a critical region of the right size for the entire family of tests.

This figure is provided on the line labeled " 40 Bonferroni." Thus if we wish to compare Hawaii's performance (among children whose parents were college graduates)with that of each of the other 40 states that participated in the state assessment, we can declare any differences greater than 11.3 points "statistically significant beyond the .05 level." The 11.3 point decision rule is conservative (for most comparisons about 50% too large), but it provides a safe and quick rule-of-thumb.

If someone is interested in comparing each state with all others there will be $\binom{41}{2}$, or 820 comparisons. To control the type I error rate in this situation requires boosting the critical region still further (to 14.0). Such a figure is given next to the label "820 Bonferroni." We are uncertain of the usefulness of including such a figure, although it still works, because anyone who is really interested in making all 820 comparisons will probably want a somewhat less conservative decision rule. Such a user of the data table would need to go back to the original standard errors and compute a more precise figure. For this particular data set, even though the standard errors are all reasonably similar, a 30 to 60 percent shrinkage in the decision rule will occur if the individual standard errors are used. Since our purpose here is not to explore alternative schemes for multiple comparisons we will not dwell on this aspect too much longer. However, depending on the circumstance, it may be profitable to choose a less conservative summary (i.e., Benjamini & Hochberg, in press). The interested reader is referred to Williams, Jones & Tukey (1994) for a much fuller exploration of the issues surrounding adjustment procedures for multiple comparisons within NAEP.

A completely revised version of the original table (from Wainer, 1995) is shown as Table 2. This table replaces the standard error columns completely with the summary standard errors. We have also reordered the rows of the table by the overall state performance[1], inserted other summary statistics for comparison, spaced the table according to the data, and indicated unusual data values. Although a description of the value of these other changes is outside the purview of this presentation, we include this complete revision as an indication of how the revised error terms fit into a broader presentation scheme.

We have learned several lessons from this tabular experiment. First, that including the standard errors of every data entry clogs the display, hindering our vision of the primary data structure. Second, that the inclusion of this additional complexity may not be especially helpful, even for the job that the standard errors are intended. Third, that a little thought about the prospective use of the data contained in the table may allow both a considerable simplification as well as increasing the ease of use. And finally fourth, that through this consideration we can provide protection for the user of the table (in this case through the explicit inclusion of the stepped-up confidence bounds for multiple compar-

---

[1] The numbers alongside each state are location aids. There is a separate locator table, alphabetical by state name, which provides these numbers for easy lookup. The idea of including an alphabetical locator table is common on maps ("Albany M6"), but has been in use for large tables for at least a century (e.g., Francis Walker used it frequently in the tables describing the growth of cities in the 1890 census.

isons) against the naive use of the standard errors. Thus the implementation demonstrated in Table 2 satisfies all three of the desiderata of effective display of error. In the next section we will explore how these lessons translate into other display formats.

Errors in Graphs

In figure 1 is a typical display that includes both data and associated error. The data are shown as large black dots, the errors are shown as vertical bars that are one standard error long in each direction. This is as typical as the tabular form of including standard errors as parenthetical additions to the data entries.

Insert Figure 1 About Here

There are analogous versions of figure 1 for other graphical formats. In figures 2 and 3 are two (now outdated) approaches to adding a standard error bar to bar charts. Of course we need not fixate on just one standard error (which under the usual Gaussian assumptions corresponds to a 68% confidence bound on the mean), we could show one-and-a-half or two standard errors(corresponding to 95% and 99% confidence bounds). In figure 4 are four versions of how this might be done. These are all traditional usages.

Insert Figures 2, 3 and 4 About Here

An aside on accurate labeling

It is important to explain exactly what the bounds depicted are bounding. In each of the cases illustrated so far they are confidence bounds on the mean of the distribution. In the past there have been clear errors of understanding on this point. As an example consider the graph in Figure 5 which was mislabeled by its producers and subsequently described erroneously by Calvin Schmid in his otherwise accurate reportage. Clearly the bounds specified are far too large to be confidence bounds on the mean. Neither could they be an accurate depiction of the distribution of years of schooling. Obviously they are showing the means with horizontal lines and using boxes and bars to denote one and two standard deviation bounds. This would only represent the actual distribution of education if that distribution was approximately Gaussian. We suspect that this is unlikely; that it is probable that at least at the turn of the century the distribution was skewed with a long straggly tail upward. How this skewness changed over time (if it did) would be an interesting question, unanswerable from this depiction.

Insert Figure 5 About Here

An obvious modification to these figures would be plotting something analogous to the Bonferroni bounds, or perhaps half a bound. In the former case anytime a point falls within such a bound from another point they are not significantly different. In the latter, any overlap of the bars sticking out could be interpreted as 'not significant'. This sort of idea has been ingeniously implemented in the NAEP charts (whose resemblance to those found on panty-hose packages, has not gone unnoticed; see figure 6) in which only states whose bounds do not overlap are different enough in performance for that difference to be thought of as consequential.

Insert Figure 6 About Here

In figure 7 is a simpler version of this same idea in which the error bars shown represent the 820 Bonferroni bounds. To compare any two states you have only to note if one state's point is outside the other's bounds. An alternative display to this (shown in Figure 8) is one in which the error bars are half of these. It appears somewhat less cluttered and comparisons require noting whether two sets of bounds overlap. We have little intuition as to which of these two options is preferred and thus will depart this aspect with

no further recommendation except to try both and see what appeals to you. Variations on this theme were explored earlier by McGill, Tukey & Larsen (1978) in their notched box plots.

Insert Figure 7 and 8 About Here

Changing the metaphor

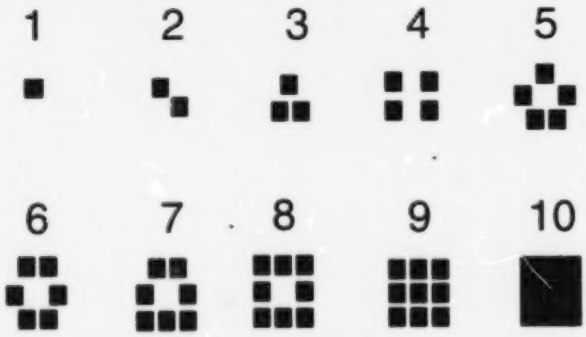
Making the physical size, and hence the visual impact, of data points proportional to their error with plotting conventions like error bars may be a bad idea. Not only does it often make the picture crowded, but it also gives greatest visual impact to the most poorly estimated points. Is this what we want to do?

For some uses such a display metaphor may be useful. For example, if we wish to draw some sort of regression line through the points, we should try to get closer to a well estimated point than to one that is more error laden. Thus we might draw less precisely estimated points with a large symbol (a longer bar) and more precisely estimated ones with a smaller symbol (a shorter bar). Then when fitting we try to make the fitted function pass through all the error bars. Often however we can implement this sort of weighted regression analytically and not have to clog the display visually with inaccurate points.

A basic premise of effective display (Bertin, 1973; Cleveland, 1985; Wainer, 1984, 1992) is to make the visual representation of the data an accurate visual metaphor. If the data increase the representation should get larger or go higher; more should indicate more. A display that violates this rule is bound to be misunderstood. How can this rule be implemented in depicting data and their associated error? At first blush it would seem to suggest that bigger error bars should be associated with bigger error. Indeed this is probably correct if the focus of the display is on the error. If however, the focus is not on the error but rather on the data themselves, perhaps we can improve the comprehensibility of the display by reversing this metaphor; by **making the size of the plotting icon proportional to the precision** and not the error. Thus data points that are more accurately estimated are drawn to make a bigger visual impression. This suggests that we make data points' size proportional to their precision (i.e., perhaps proportional to se^{-1} if we wish to focus on the precision or proportional to se^{-2} if we wish instead to focus on the sample size represented). In this way our attention will be drawn to the more accurate points; very inaccurate points would be hardly visible and so have little impact.

As an example consider the hypothetical data shown in Figure 9 There are five points shown with their standard errors. The graph suggests that the relationship between X and Y is nonlinear with an ogive shape. The same data are shown in Figure 10 in which

the data points' area are shown proportional to their precision (you must look carefully for point 4). In this depiction the linearity of the data is one's principal perception; such a perception is supported by formal statistical testing.[2] Perhaps we can generalize this idea through the use of something akin to horizontal 'precision bars' around each point. These bars could be proportional to the (Fisherian) information contained each point. This idea is merely the visual counterpart to what is commonly done already in many analytic estimation procedures, when we sometimes examine the diagonal elements of the Hessian (the standard errors of the parameters) and sometimes the inverse Hessian (the information matrix). Another alternative might be to use a plotting icon [like Bachi's (1978) "Graphical Rational Pattern" shown below] to reflect precision.



Hoaglin and Tukey's (1985) confidence-aperture plots are another example of a plot in which the least accurately estimated points are depicted with mostly empty space (see figure 11).

The worth of such schemes will require some visual experimentation with formats, some experience to overcome the older conventions, and the careful consideration of the questions that the display is meant to answer.

Insert Figures 9, 10, and 11 About Here

Sloppy (dirty) graphics

Another graphical way to remind the viewer of the fallibility of the data in a display is to draw and label the display only as well as the accuracy of the data allow. This is exactly analogous to the too often ignored practice of writing down numerical values to only the number of digits that are significant (for arguments and illustrations of the effi-

---

[2] A likelihood ratio test shows no significant improvement in fit for models allowing deviation from linearity.

cacy and importance of rounding see Ehrenberg, 1977; Walker & Durost, 1936; Wainer, 1993, 1995).

One possible approach might be to calibrate and label the axes in such a way so as to limit the accuracy of the inferences to only what is justified by the data. An example of such a display is shown in figure 12.

Insert Figure 12 About Here

While an asset of such a display is that it doesn't allow incorrectly precise statements, a possible drawback is that it doesn't allow imprecisely correct statements. This provokes the image of Damon Runyun, when he said, "The race is not always to the swift, nor the battle to the strong, but that's the way to lay your bets."[3] Thus even though we may not have good estimates it may sometimes be important for the graph to allow us to extract the best guess available. I have only a little sympathy for this view.[4]

An allied approach might be to accurately label and calibrate the axes but remove distinct boundaries from the graphical representation. For example, in figure 13 we show the performance of five ethnic groups on one portion of the 1992 NAEP mathematics assessment. Through the utilization of graduated shading around the actual mean performance we make it impossible to unambiguously choose a single number to characterize that ethnic group's performance. The location at which the shading begins to thin is determined by both the mean level and the estimated standard error. Groups, like American Indians, with large standard errors have the shading begin well below the mean and it changes very gradually. Ethnic groups, like Whites, that have small standard errors are characterized by a much steeper gradient in the shading algorithm.[5] This approach, is a continuous analog[6] to the more traditional methods shown in figure 4

---

[3]Perhaps, specializing John Tukey's well known aphorism, this corresponds to "If it's worth displaying, it's worth displaying badly."

[4]I have often witnessed requests for a single number to characterize a complex situation. Some then argue that "a single number is inappropriate and we ought not provide it." Others cynically contend that 'if we don't give them a number they'll calculate one of their own. Let's at least provide the best single number we can." I have some sympathy with both views. Perhaps by providing graphical answers like that in figure 12 we can provide the qualitative answer without allowing inappropriate precision.

[5]Figure 13 is little more than an initial attempt. I fully expect that it can be improved upon with experience. The rule used to generate it was to use black at the mean and then graduate the shading exponentially to white 2.5 se above and below the mean.

[6]This is probably more properly called a Bayesian approach since it depicts a posterior distribution rather than a confidence region.

Insert Figure 13 About Here

Summary

In this essay we have described and illustrated some suggestions for alternative methods to depict error. The principal goal was to make it easier for inferences about data to include the effects of the precision with which those data were measured. Sometimes (with tabular presentation) this involved calculating and displaying the appropriate error term for the most likely uses of the data. In other situations (like the fade-away chart in figure 12) it meant making the precision an integral part of the display. I'm sure that the notion of using shading to characterize error is a long way from being original with me (my former colleague, Albert Biderman, suggested it to me over coffee a decade ago). I am certain that the primary reason we have not seen it used more widely in practice is its difficult implementation. Modern computer graphic software has removed this impediment. This apparently novel statistical suggestion is old hat to cartographers. Petermann (1851) used continuous tone implemented by lithographic crayon shading to show variations in the population of Scotland. But even this use was presaged by Quetelet (1832) who produced some fuzzy social data maps showing the distribution of crime (see Figure 14, reproduced from Robinson, 1982). More recently, MacEachren (1992, p. 14; 1994, p. 81) used this method effectively to display the distribution of risk surrounding a nuclear power plant. He also suggests schemes such as defocusing images to convey uncertainty.

Insert Figure 14 About Here

Dynamic possibilities for the representation of uncertainty have not been discussed, although we believe that such schemes as blinking data points in which the proportion of 'on' to 'off' time is proportional to the points' precision, or a visual analog of multiple imputation in which the location of a data point changes as you watch it, are promising possibilities.

The suggestions contained here are meant only as a beginning. The assessment of their usefulness awaits broader application, discussion and experimentation. Paraphrasing Karl Pearson's final words in the *Grammar of Science*, it is only by "daring to display our ignorance" will our perceptions of scientific progress stay in reasonably close proximity to that progress.

References

Bachi, R. (1978). Proposals for the development of selected graphical methods. In *Graphic presentation of statistical information: Papers presented at the 136th annual meeting of the American Statistical Association*. U.S. Dept. of Commerce, Bureau of the Census.

Benjamini, Y. & Hochberg, Y. (in press). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, B.*

Bertin, J. (1973). *Semiologie Graphique*. The Hague: Mouton-Gautier. 2nd Ed. (English translation done by William Berg & Howard Wainer and published as *Semiology of Graphics,* Madison, Wisconsin: University of Wisconsin Press, 1983.)

Bonferroni, C. E. (1936). Il Calcolo delle assicurazioni su gruppi di teste. In *Studii in onore del prof. S. O. Carboni.* Roma.

Cleveland, W. S. (1985). *The elements of graphing data.* Wadsworth: Monterey, CA.

Ehrenberg, A. S. C. (1977). Rudiments of numeracy. *Journal of the Royal Statistical Society, Series A, 140,* 277-297.

Hoaglin, D. C. & Tukey, J. W. (1985) Checking the shape of discrete distributions. Chapter 9 in David C. Hoaglin, Frederick Mosteller & John W. Tukey (eds.) *Exploring data tables, trends and shapes.* pp. 345-416. New York: Wiley.

MacEachren, A. M. (1992). Visualizing uncertain information. *Cartographic perspectives, 13,* 10-19.

MacEachren, A. M. (1994). Visualization quality and the representation of uncertainty. Chapter 4 in *Some Truth with Maps: A Primer on symbolization and design.* Washington, DC: Association of American Cartographers.

McGill, R., Tukey, J.W., & Larsen, W. (1978). Variations of box plots. *The American Statistician, 32*(1), 12-16.

Miller, R. G. (1966). *Simultaneous statistical inference.* New York: McGraw-Hill.

Mosteller, F., & Tukey, J.W. (1968). "Data analysis, Including Statistics." In G. Lindzey & E. Aronson (eds.) *The Handbook of Social Psychology, Vol. 2,* 80-203. Reading, MA: Addison-Wesley.

Petermann, A. (1852). Distribution of the Population. *Census of Great Britain, 1851.* Lithograph, crayon shading in the British Library.

Robinson, A. H. (1982). *Early thematic mapping in the history of cartography*. Chicago: University of Chicago Press.

Quetelet, A. (1832). Recherches sur le penchant au crime aux différens ages, *Nouveaux Mémoires de l'Académie Royales des Sciences et Belle-Lettres de Bruxelles, 7,* 1-87.

Schmid, C. F. (1983). *Statistical graphics: Design Principles and Practices*. New York: John Wiley & Son.

Wainer, H. (1984). How to display data badly. *The American Statistician, 38,* 137-147.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher, 21,* 14-23.

Wainer, H. (1993). Tabular presentation. *Chance, 6(3),* 52-56.

Wainer, H. (1995). *A study of display methods for NAEP results: I. Tables,* ETS Technical Report 95-xx. Princeton, NJ: Educational Testing Service.

Walker, H. M. & Durost, W. N. (1936). *Statistical Tables: Their structure and use.* New York: Bureau of Publications, Teachers College, Columbia University.

Williams, V. S. L, Jones, L. V. & Tukey, J. W. (1994). *Controlling error in multiple comparisons with special attention to the National Assessment of Educational Progress.* Technical Report # 33. Research Triangle Park, NC: National Institute of Statistical Sciences.

Table 1

Average Mathematics Proficiency by Parents' Highest Level of Education Grade 8 - 1992

PUBLIC SCHOOLS	Graduated College		Some Education After High School		Graduated High School		Did Not Finish High School		I don't Know	
	θ	se	θ	se	θ	se	θ	se	θ	se
NATION	**279**	**1.4**	**270**	**1.2**	**256**	**1.4**	**248**	**1.8**	**251**	**1.7**
STATES										
Alabama	261	2.5	258	2.0	244	1.8	239	2.0	237	2.9
Arizona	277	1.5	270	1.5	256	1.6	245	2.5	248	2.7
Arkansas	264	1.9	264	1.7	248	1.6	246	2.4	245	2.7
California	275	2.0	266	2.1	251	2.1	241	2.2	240	2.9
Colorado	282	1.3	276	1.6	260	1.5	250	2.4	252	2.6
Connecticut	288	1.0	272	1.8	260	1.8	245	3.3	251	2.4
Delaware	274	1.3	268	2.3	251	1.7	248	4.0	248	3.4
District of Columbia	244	1.7	240	1.9	224	1.6	225	3.2	229	2.2
Florida	268	1.9	266	1.9	251	1.8	244	2.7	244	3.2
Georgia	271	2.1	264	1.7	250	1.3	244	2.2	245	2.6
Hawaii	267	1.5	266	1.9	246	1.8	242	3.5	246	2.1
Idaho	281	0.9	278	1.3	268	1.4	254	2.3	254	2.8
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

Error terms for comparisons					
Max Std error of difference	3.5	3.4	3.5	6.4	6.4
40 Bonferroni	11.3	11.0	11.3	20.7	20.7
820 Bonferroni	14.0	13.6	14.0	25.6	25.6

Source: Abstracted from Table 2.12 in
The 1992 NAEP Trial State Assessment

Table 2

Average Mathematics Proficiency by Parents' Highest Level of Education
Grade 8 - 1992

	PUBLIC SCHOOLS	Graduated College	Some Education After High School	Graduated High School	Did Not Finish High School	I Don't Know	Mean
	NATION	**279**	**270**	**256**	**248**	**251**	**267**
	STATES						
1	Iowa	291	285	273	262	266	**283**
2	North Dakota	289	283	271	259	272	**283**
3	Minnesota	290	284	270	256	268	**282**
4	Maine	288	281	267	259	266	**278**
5	Wisconsin	287	282	270	254	255	**278**
6	New Hampshire	287	280	267	259	262	**278**
7	Nebraska	287	280	267	**247**	256	**277**
8	Idaho	281	278	268	254	254	**274**
9	Wyoming	281	278	266	258	260	**274**
10	Utah	280	278	258	254	258	**274**
11	Connecticut	288	272	260	**245**	251	**273**
12	Colorado	282	276	260	250	252	**272**
13	Massachusetts	284	272	261	248	**248**	**272**
14	New Jersey	283	275	259	253	250	**271**
15	Pennsylvania	282	274	262	252	252	**271**
16	Missouri	280	275	264	254	252	**271**
17	Indiana	283	275	260	250	249	**269**
18	Ohio	279	272	260	243	249	**268**
19	Oklahoma	277	272	257	254	251	**267**
20	Virginia	282	270	252	248	251	**267**
21	Michigan	277	271	257	249	248	**267**
22	New York	277	271	256	243	**240**	**265**
23	Rhode Island	276	271	256	244	**239**	**265**
24	Arizona	277	270	256	245	248	**264**
25	Maryland	278	266	250	240	245	**264**
26	Texas	281	272	253	247	244	**264**
27	Delaware	274	268	251	248	248	**262**
28	Kentucky	278	267	254	246	242	**261**
29	California	275	266	251	241	240	**260**
30	South Carolina	272	268	248	248	247	**260**
31	Florida	268	266	251	244	244	**259**
32	Georgia	271	264	250	244	245	**259**
33	New Mexico	272	264	249	244	245	**259**
34	Tennessee	267	265	251	245	243	**258**
35	West Virginia	270	**269**	251	244	239	**258**
36	North Carolina	271	265	246	240	240	**258**
37	Hawaii	267	266	246	242	246	**257**
38	Arkansas	264	264	248	**246**	245	**256**
39	Alabama	261	258	244	239	237	**251**
40	Louisiana	256	259	242	237	236	**249**
41	Mississippi	254	256	239	234	231	**246**
	Means	**279**	**270**	**256**	**248**	**251**	**267**
	TERRITORIES						
42	Guam	246	244	229	224	226	**238**
43	District of Columbia	244	240	224	225	229	**234**
44	Virgin Islands	224	232	221	219	217	**222**
	Error terms for comparisons						
	Max Std error of diff	3.5	3.4	3.5	6.4	6.4	3.5
	40 Bonferroni	11.3	11.0	11.3	20.7	20.7	11.3
	820 Bonferroni	14.0	13.6	14.0	25.6	25.6	14.0

Figure Captions

Figure 1. A typical depiction of data including both their level and error. The large dots locate the position of the mean, the bars depict one standard error of the mean in each direction.

Figure 2. Figure 10-1 from Schmid (1983, p. 192) showing a traditional method of depicting error on a bar chart using an error rectangle on the side of the bar. Reprinted with permission.

Figure 3. Figure 10-2 from Schmid (1983, p. 193) showing a traditional method of depicting error on a bar chart using an error rectangle superimposed over the end of the bar. Reprinted with permission.

Figure 4. Figure 10-3 from Schmid (1983, p. 195) showing four methods of depicting confidence bounds on a bar chart using various figurations superimposed over the end of the bar. Reprinted with permission.

Figure 5. Figure 10-7 from Schmid 1983, p. 199) showing how incorrect labeling can lead to wildly inaccurate inferences. It reports confidence intervals on the mean, but in fact merely depicts the mean and error bars showing one and 2 standard deviations in each direction. Reprinted with permission.

Figure 6. Figure 1.6 from the *National Compendium for the NAEP 1992 Mathematics Assessment for the Nation and the States* (p. 40). A "Pantyhose chart" that includes the Bonferroni expanded standard error of the mean within its bounds so that the user can scan down any column (state) and instantaneously decide which states are significantly different in their performance from that state.

Figure 7. A difference between two states is statistically significant (at the .05 level) if one state's data point is outside the other state's bounds.

Figure 8. A difference between two states is statistically significant (at the .05 level) if one state's bounds overlap the other state's bounds.

Figure 9. A plot of hypothetical data that displays the data and their standard errors as vertical bars. The graph suggests a nonlinear relationship between X and Y.

Figure 10. A plot of hypothetical data that displays the data sized proportional to their precision. The graph suggests a linear relationship between X and Y.

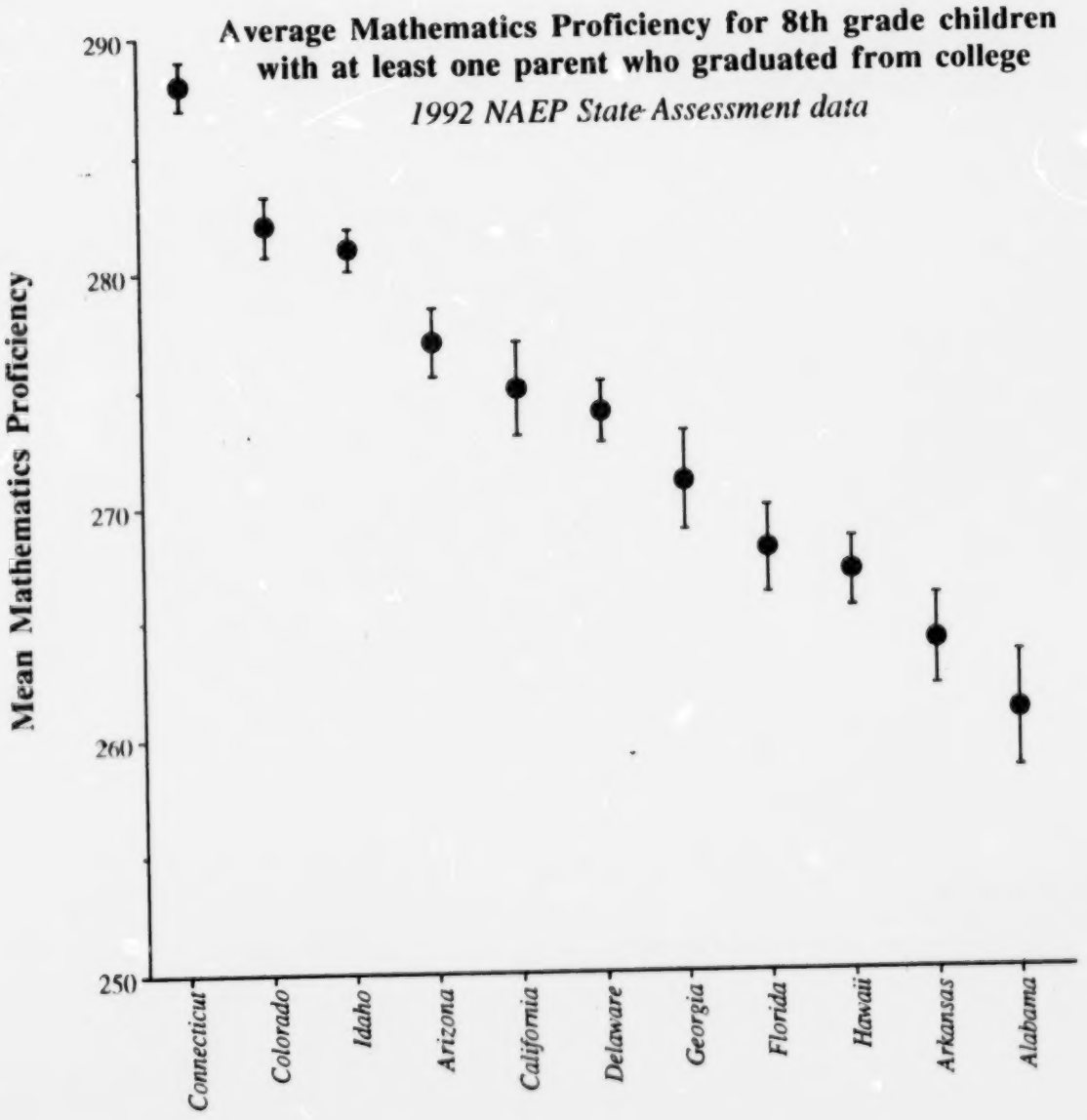
Figure 11. A confidence-aperture plot from Hoaglin & Tukey (1985, figure 9-14, p. 386) in which the amount of open space reflects the uncertainty. Reprinted with permission.

Figure 12. Displaying data of uncertain precision with the accuracy they deserve through the use of appropriately calibrated axes.

Figure 13. Using graduated shading can make it possible to automatically build the perception of variability into what used to be a bar chart .

Figure 14. Graduated shading was used by Quetelet in 1831 to show the distribution crimes against property (left side) and crimes against persons (right side). Reprinted with permission from Robinson (1982).

Figure 1


Average Mathematics Proficiency for 8th grade children with at least one parent who graduated from college
1992 NAEP State Assessment data
Mean Mathematics Proficiency
290
280
270
260
250
Connecticut
Colorado
Idaho
Arizona
California
Delaware
Georgia
Florida
Hawaii
Arkansas
Alabama

Unemployment Among High School Graduates 16 to 24 Years Old, by Race: October 1975

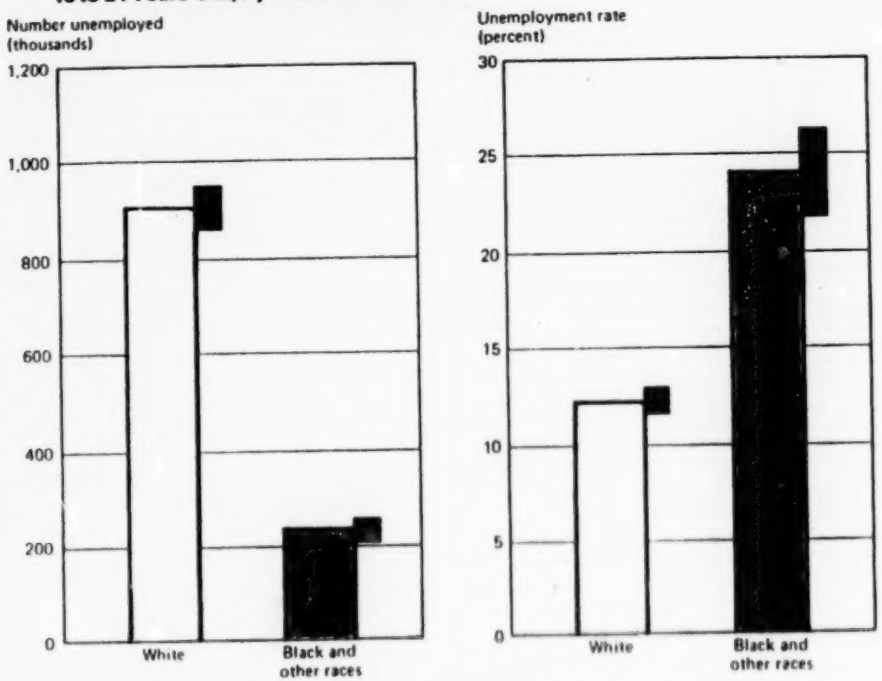
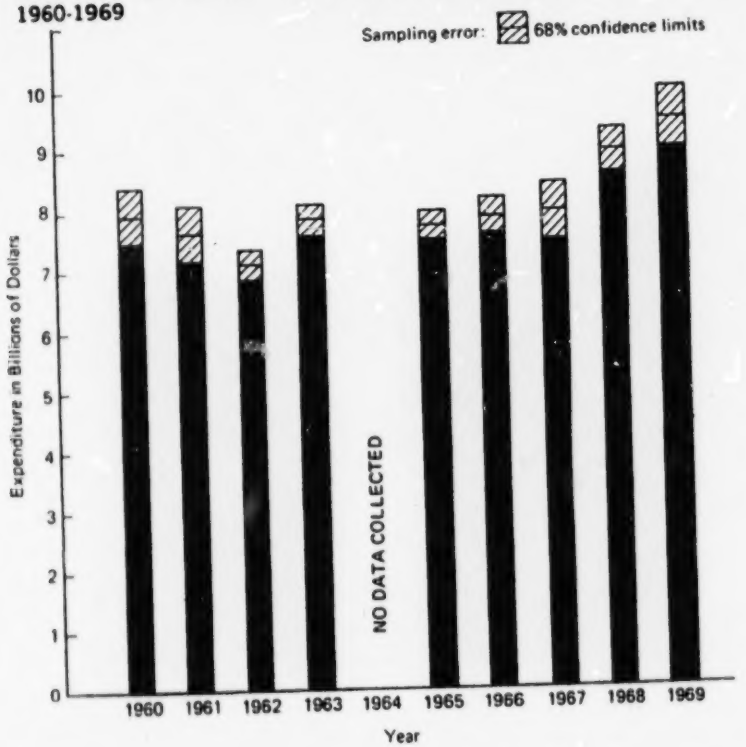


Figure 10-1. The data depicted by this chart are based on a sample survey. The amount of sampling variability of the data is indicated by the small rectangles. The height of the rectangles covers a range of values of one standard error above and one standard error below the reported value. (From Office of Federal Statistical Policy and Standards, U.S. Bureau of the Census, Social Indicators, 1976, Washington, D.C.: Government Printing Office, 1977, p. XXVII.)

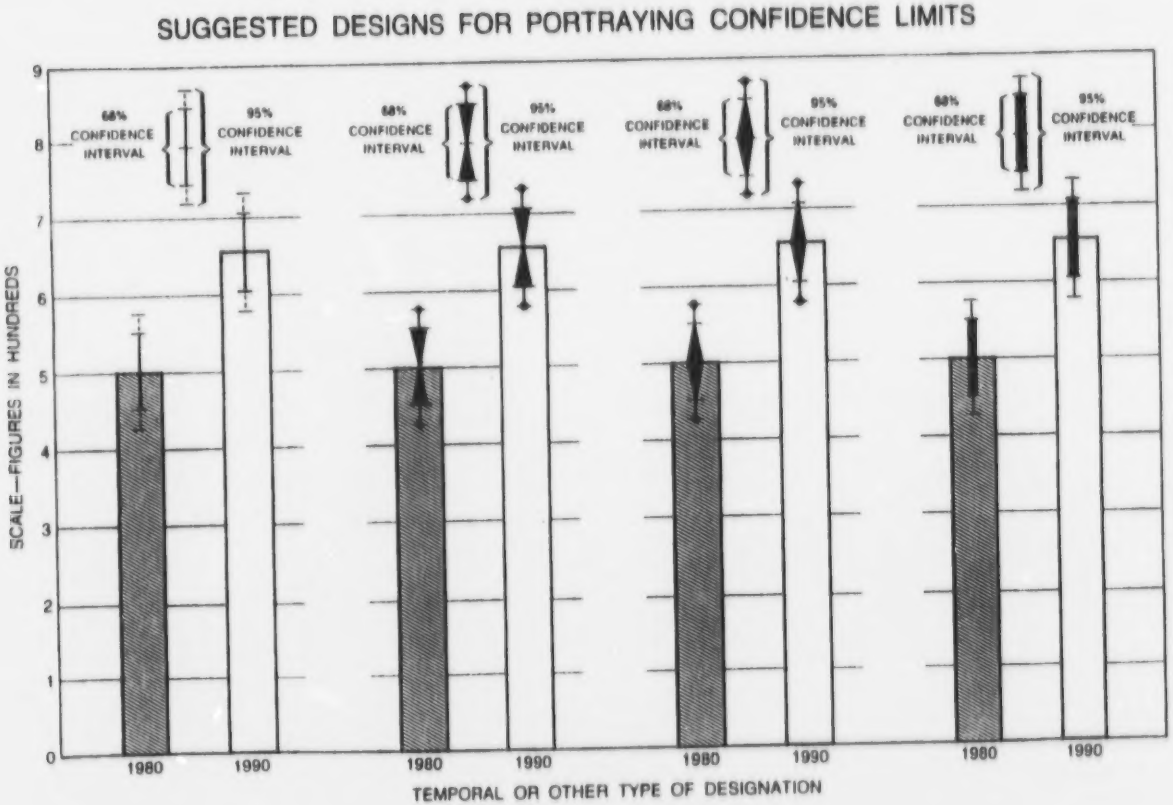
Figure 10-2. This chart, like the previous one, shows the range of 68-percent confidence limits of the data, not by separate rectangles but by cross-hatching. (From Maria E. Gonzalez et al., Standards For Discussion and Presentation of Errors in Data, Technical Paper 32, United States Bureau of the Census, Washington, D.C.: Government Printing Office, 1974, p. v-3.)

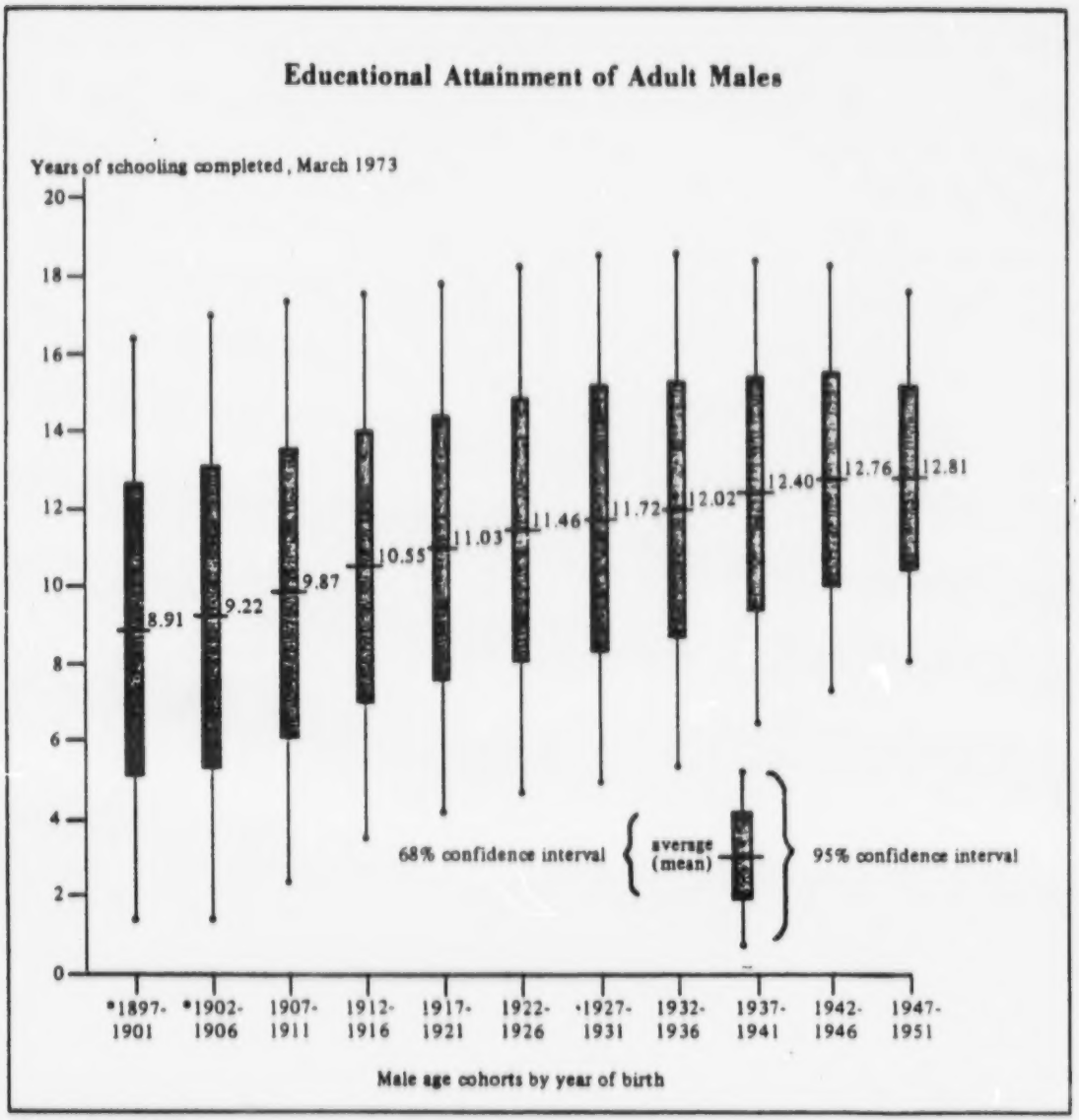
Estimates of Expenditures by Owner-Occupants for Alterations and Repairs to Their 1-to-4-Unit Residential Properties as Measured by Quarterly Surveys, 1960-1969



Note: Estimates for 1960 and 1961 are revised to adjust for bias in originally published figures, which we were able to measure and correct.

Figure 10-3. Additional designs for portraying confidence limits for simple column and bar charts. Certain features based on suggestions of Albert Biderman and staff, Bureau of Social Science Research, Washington, D.C.





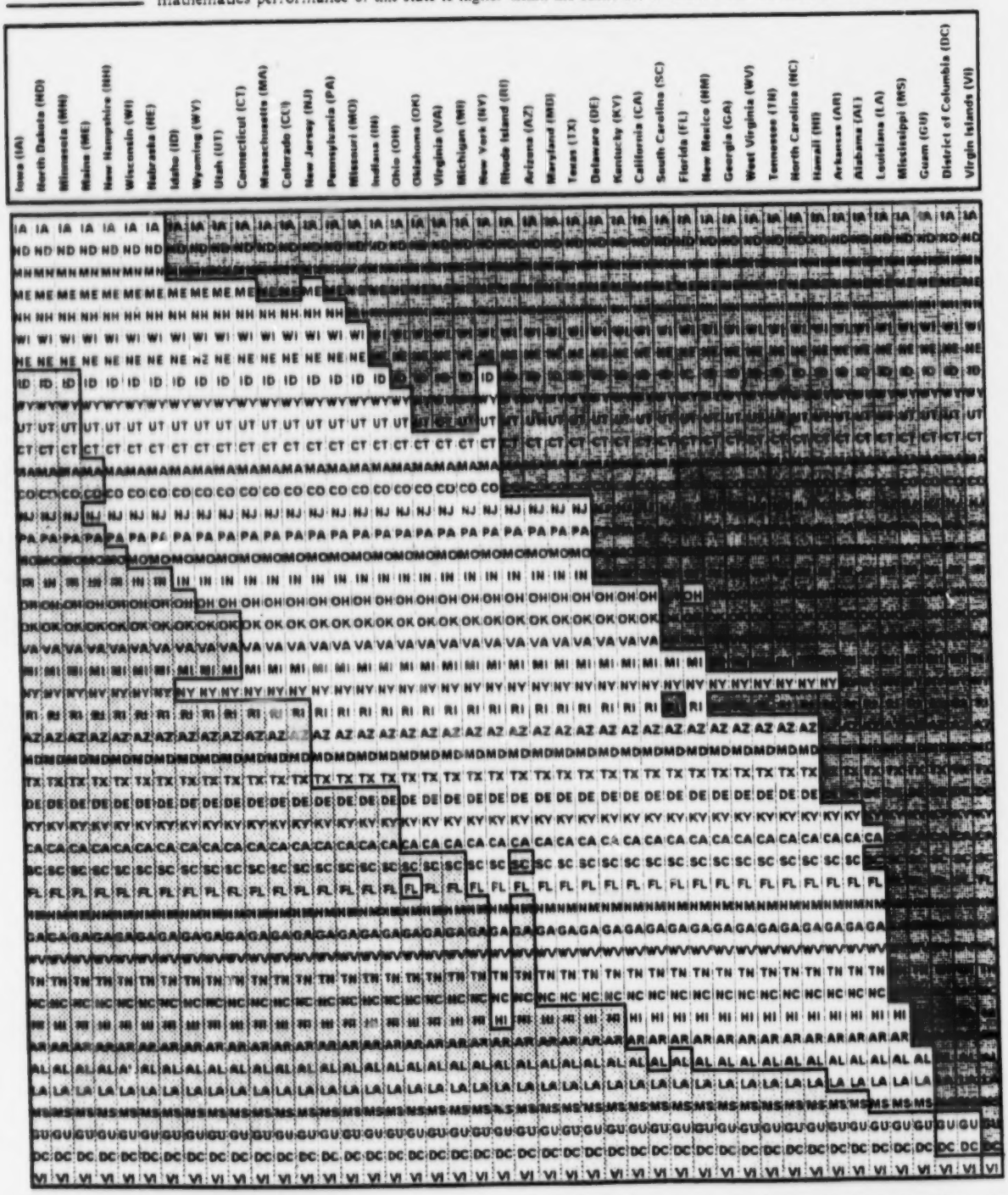
*From 1962 OCG Survey
Source of Data: Bureau of the Census and special tabulations by Hauser & Featherstone

Figure 10-7. In this chart, the portrayal of sampling variability of the basic data has been incorporated as an integral part of the chart design. The vertical bars and lines indicate specifically 68- and 95-percent confidence intervals, while short horizontal lines and figures show estimates for each time period. (From Mary A. Golladay, The Condition of Education, Part 1, HEW National Center for Education Statistics, Washington, D.C.: Government Printing Office, 1977, p. 104.)

FIGURE 1.6 | **Comparisons of Overall Mathematics Average Proficiency 1992 Grade 8**

THE NATION'S REPORT CARD

INSTRUCTIONS: Read *down* the column directly under a state name listed in the heading at the top of the chart. Match the shading intensity surrounding a state postal abbreviation to the key below to determine whether the average mathematics performance of this state is higher than, the same as, or lower than the state in the column heading.



- State has statistically significantly higher average proficiency than the state listed at the top of the chart.
- No statistically significant difference from the state listed at the top of the chart.
- State has statistically significantly lower average proficiency than the state listed at the top of the chart.

The between state comparisons take into account sampling and measurement error and that each state is being compared with every other state. Significance is determined by an application of the Bonferroni procedure based on 946 comparisons by comparing the difference between the two means with four times the square root of the sum of the squared standard errors.

Average Mathematics Proficiency for 8th grade children with at least one parent who graduated from college

1992 NAEP State Assessment data

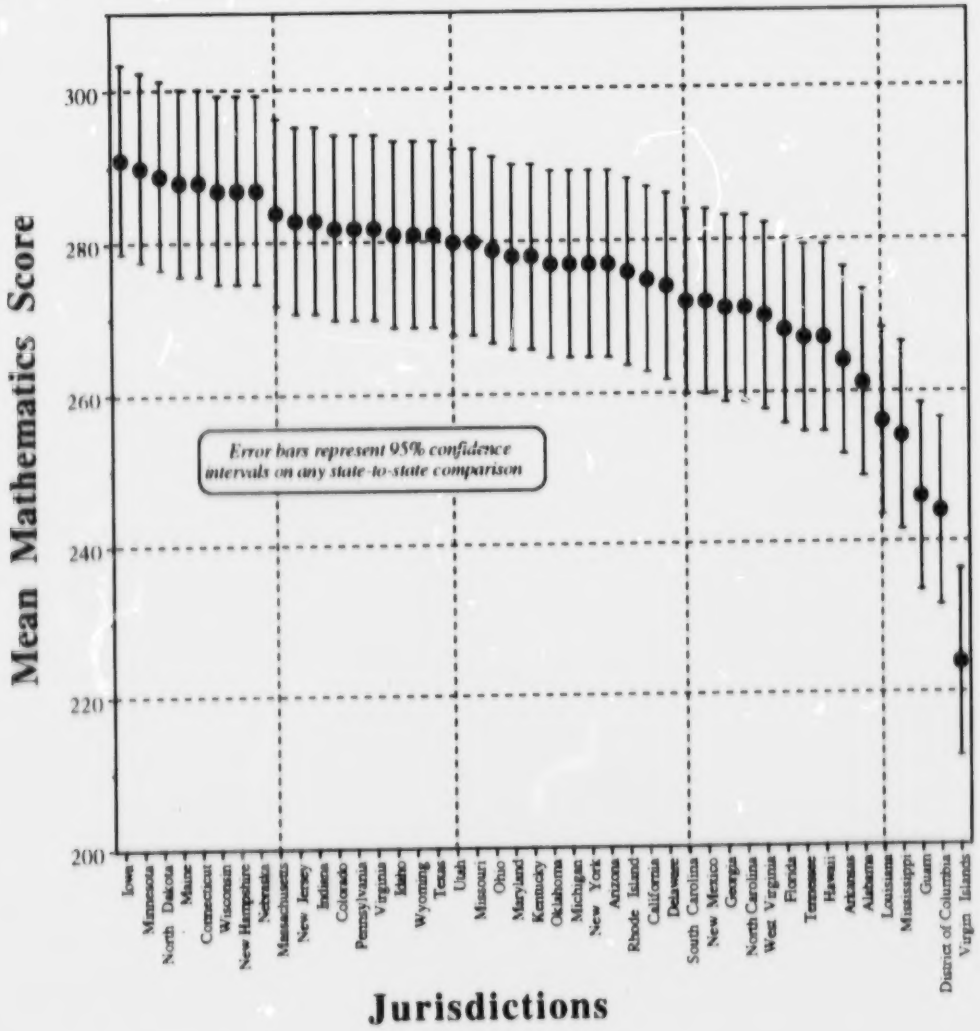


Figure 7

Average mathematics proficiency for 8th grade children with at least one parent who graduated from college
1992 NAEP State Assessment data

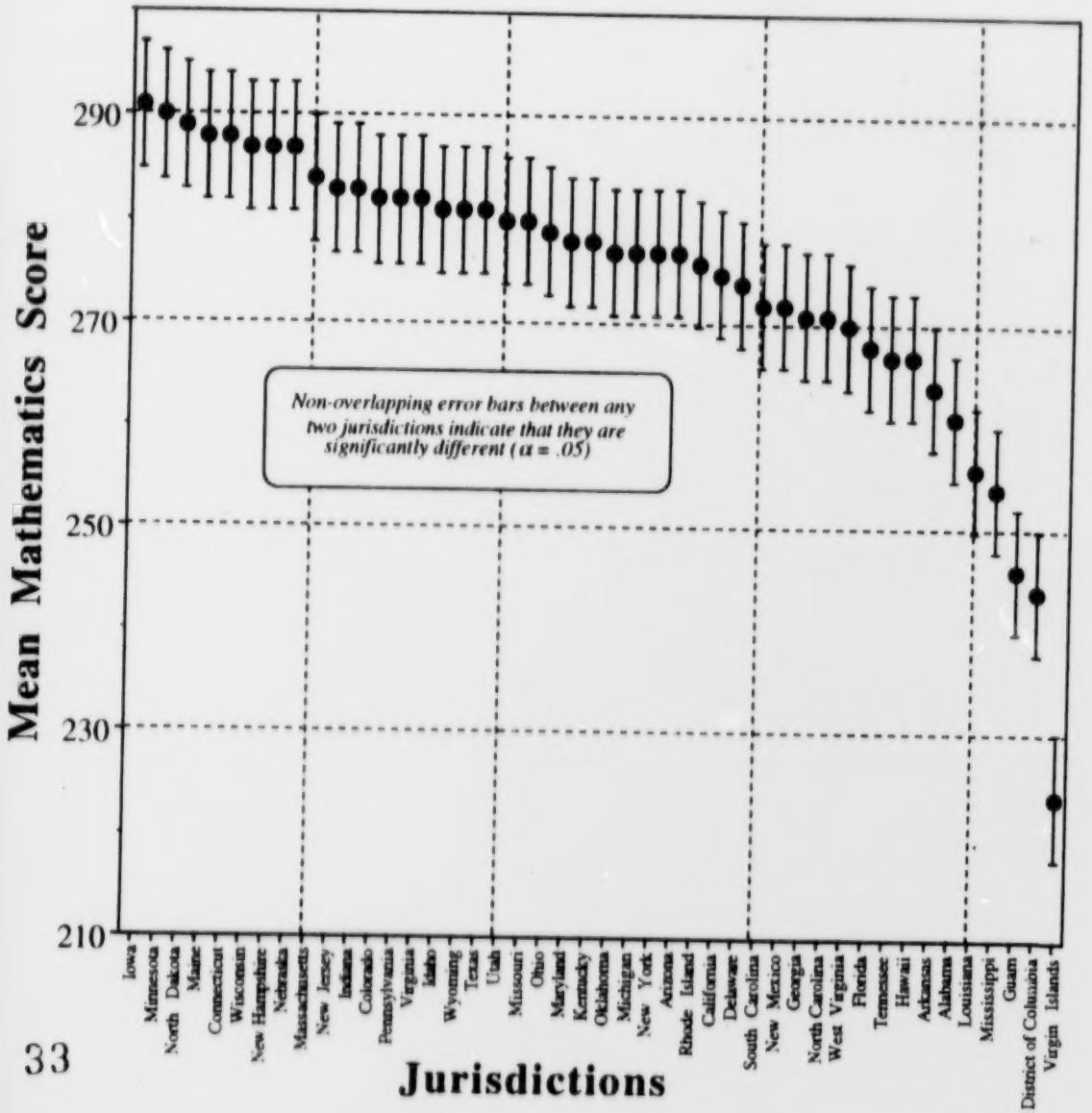
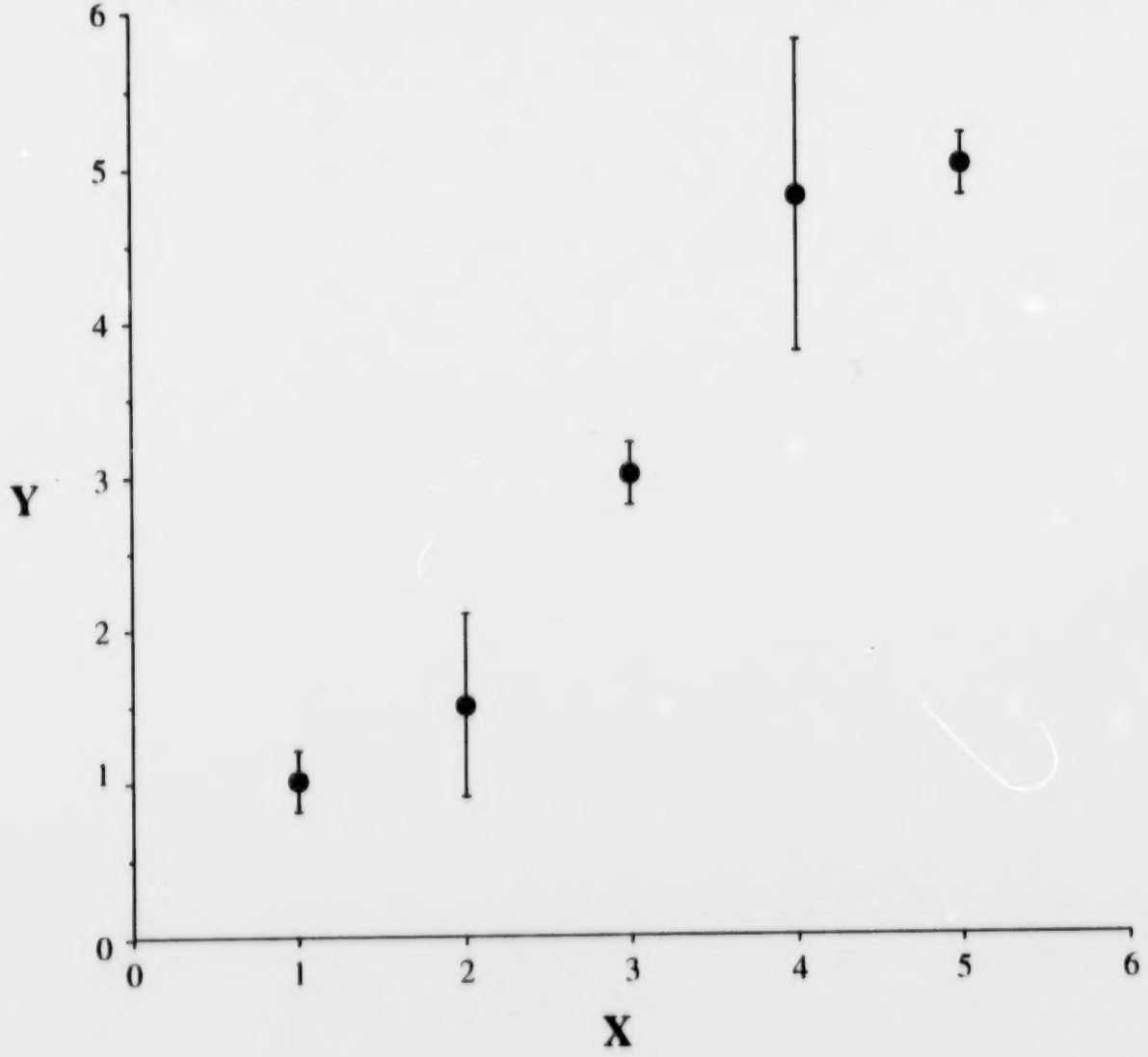


Figure 8

Hypothetical data showing error bars



35

36

Figure 9

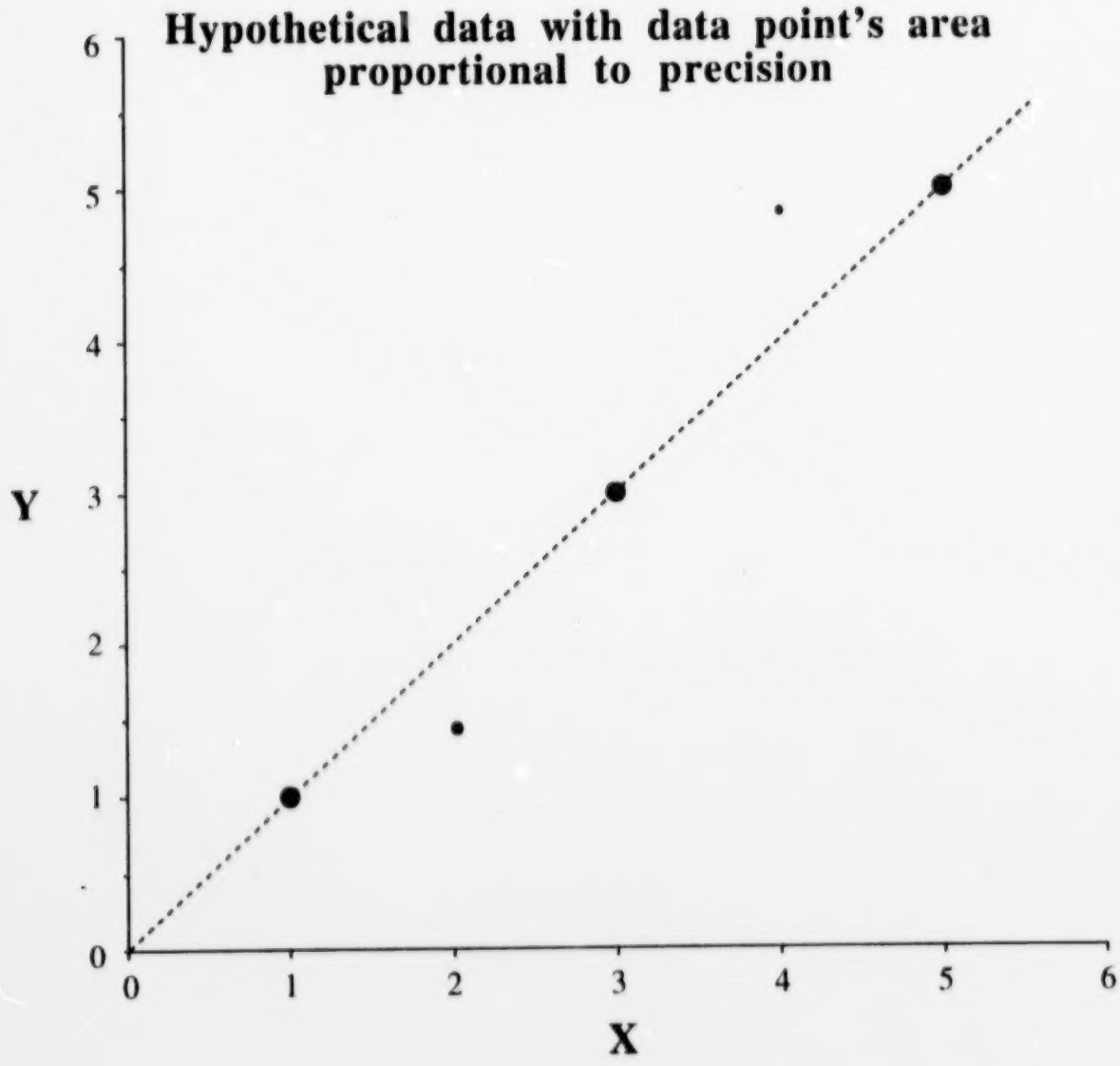

Hypothetical data with data point's area proportional to precision
Y
X
0
1
2
3
4
5
6

Figure 10

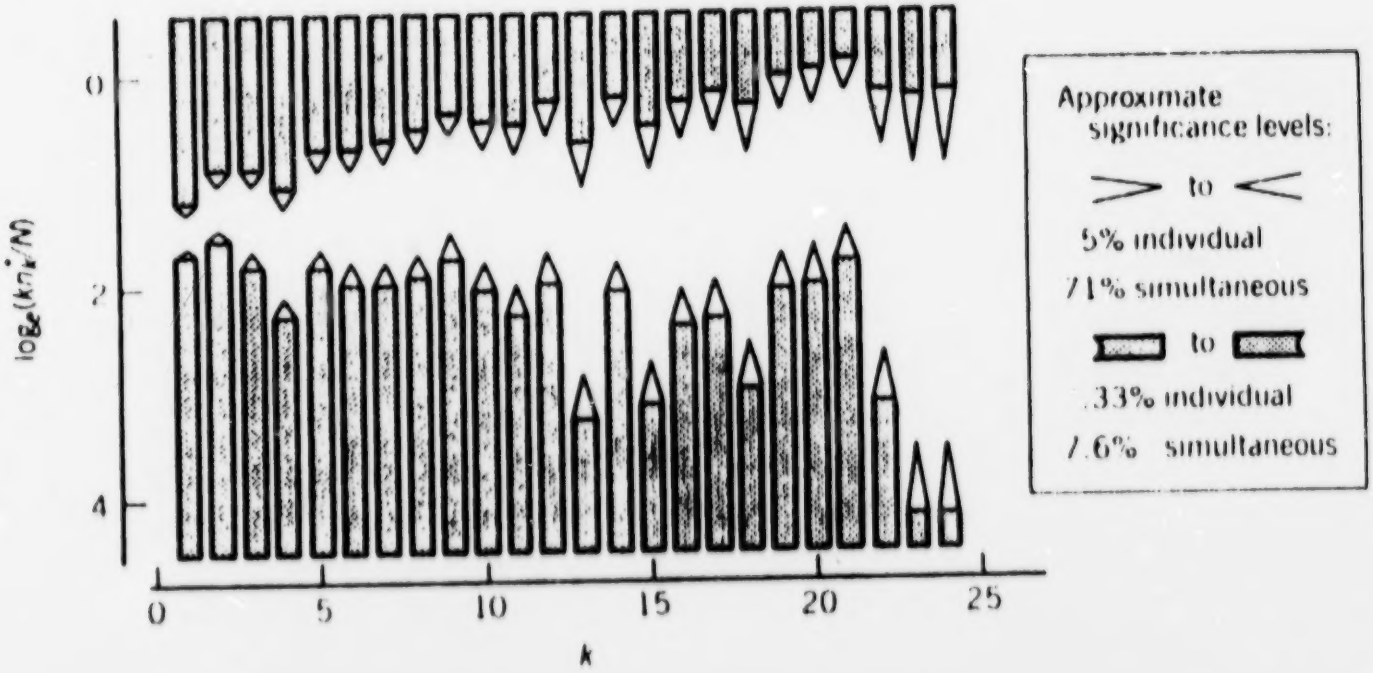


Figure 9-14. Confidence-aperture plot for the butterfly species data.

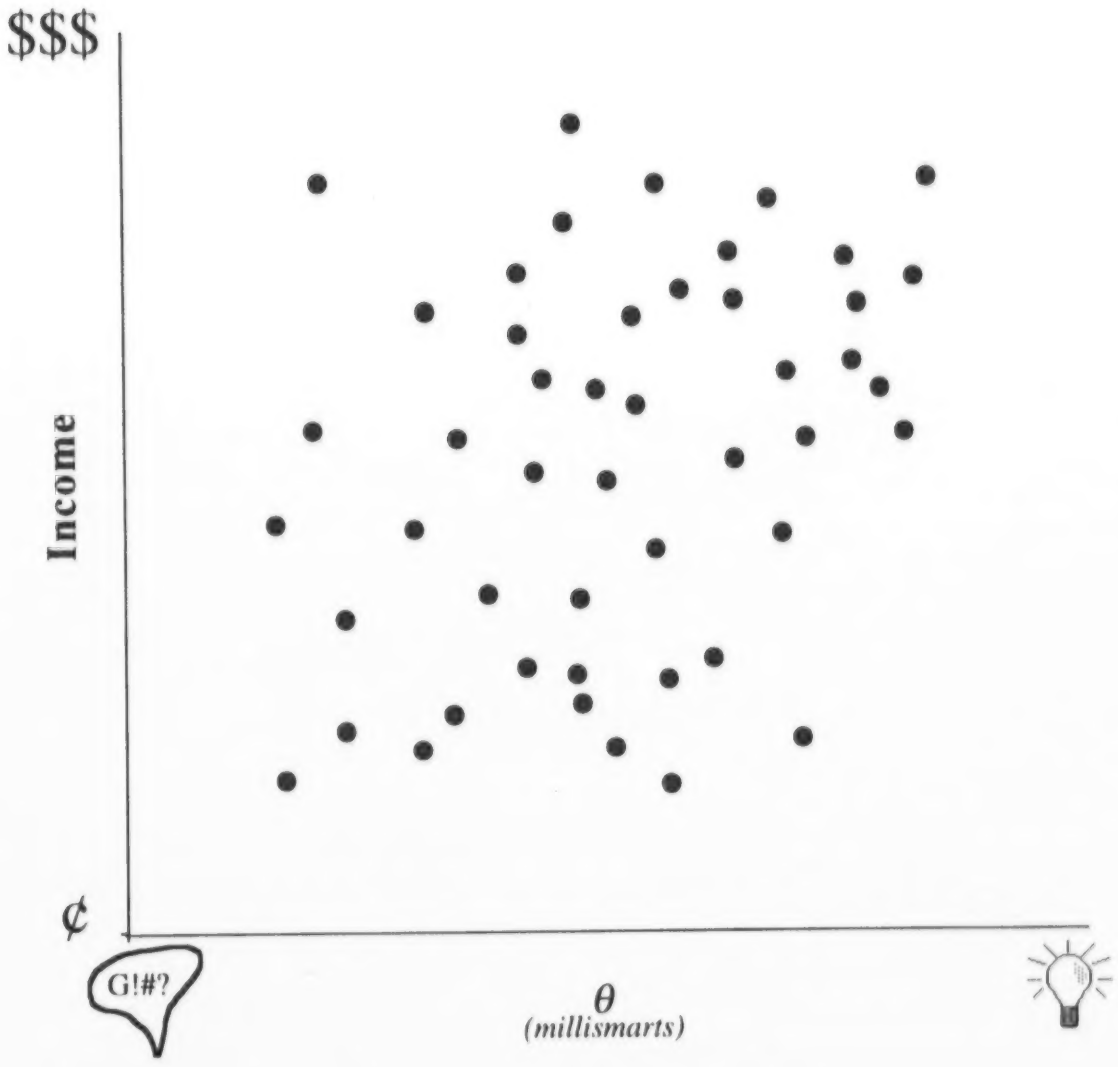


Figure 12

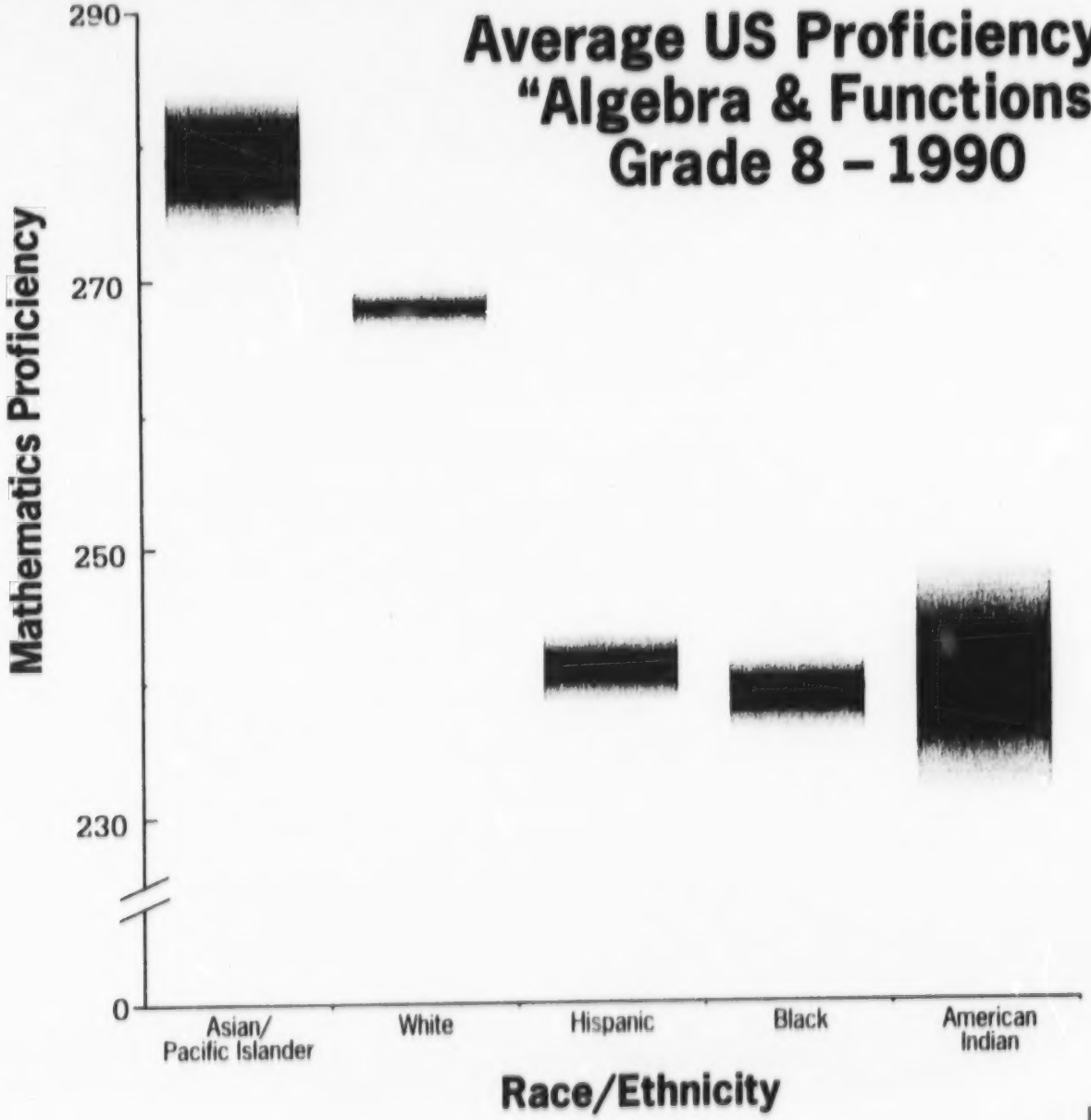

Average US Proficiency in "Algebra & Functions" Grade 8 – 1990
Mathematics Proficiency
290
270
250
230
0
Asian/ Pacific Islander
White
Hispanic
Black
American Indian
Race/Ethnicity

Figure 13

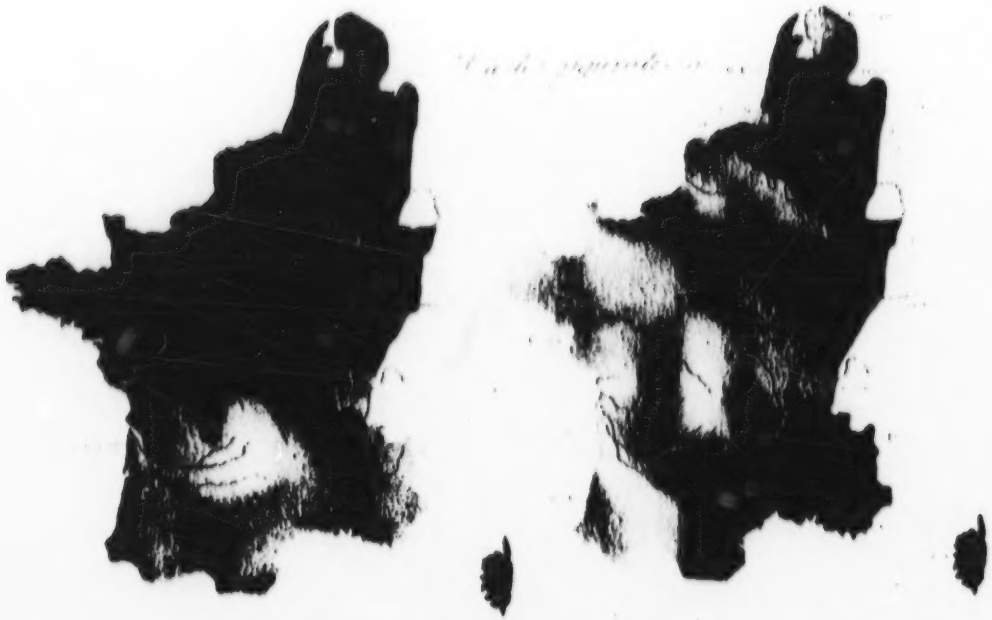


Figure 77 Quetelet's 1831 plate 2 "Cartes figuratives—Crimes contre les propriétés—Contre les personnes." Original 335 × 230 mm. Lithograph, crayon shading. (Courtesy of the Center for Research Libraries, Chicago.)

Figure 14